Determining the Number of Regimes in a Threshold Autoregressive Model Using Smooth Transition Autoregressions*

Birgit Strikholm †
Department of Economic Statistics
Stockholm School of Economics

Timo Teräsvirta‡
Department of Economic Statistics
Stockholm School of Economics

SSE/EFI Working Paper Series in Economics and Finance, No 578
January 2005

Abstract

In this paper we propose a method for determining the number of regimes in threshold autoregressive models using smooth transition autoregression as a tool. As the smooth transition model is just an approximation to the threshold autoregressive one, no asymptotic properties are claimed for the proposed method. Tests available for testing the adequacy of a smooth transition autoregressive model are applied sequentially to determine the number of regimes. A simulation study is performed in order to find out the finite-sample properties of the procedure and to compare it with two other procedures available in the literature. We find that our method works reasonably well for both single and multiple threshold models.

Key words: Model specification, model selection criterion, nonlinear modelling, sequential testing, switching regression.

JEL Classification Code: C22, C51

*This research has been supported by Jan Wallander's and Tom Hedelius's Foundation, Grant No. J02-35. A part of the the work was carried out when the authors were visiting the Department of Economics, University of California, San Diego, whose kind hospitality is gratefully acknowledged. We are grateful to Jesús Gonzalo and Jean-Yves Pitarakis who gracefully allowed us to use their GAUSS code written for selecting the number of regimes using model selection criteria. Material from this paper has been presented at the EC^2 meeting, Bologna, December 2002, and the Nordic Econometric Meeting, Bergen, May 2003. We wish to thank Pentti Saikkonen and Heather Anderson for useful comments but retain the responsibility for any errors and shortcomings in this work.

†Department of Economic Statistics, Stockholm School of Economics, Box 6501, SE-113 83 Stockholm, Sweden, email: Birgit.Strikholm@hhs.se

‡Department of Economic Statistics, Stockholm School of Economics, Box 6501, SE-113 83 Stockholm, Sweden, email: Timo.Terasvirta@hhs.se

1 Introduction

The switching regression model (Quandt, 1958) and its univariate counterpart, the threshold autoregressive (TAR) model (Tong, 1978) are popular nonlinear models. The TAR model in particular has generated a wide range of papers covering both theoretical and empirical issues. An overview can be found in Tong (1990); see also, for example, Caner and Hansen (2001), Hansen (1996, 1999a, 2000), Kapetanios (2003), Koop and Potter (1999), Medeiros et al. (2002), among others.

In most economic applications of the TAR model, economic theory is not specific about the complete structure of the model. In particular, most often the number of regimes in the model cannot be assumed known *a priori*. Furthermore, the switch variable or, in the TAR case, the delay determining the threshold variable is often unknown as well. Some work exists on how to select the number of regimes in TAR models. Tsay (1989) suggested a graphical approach for locating the values of thresholds. He used scatterplots of standardized predictive residuals (in arranged autoregression) and recursive t-ratios of an AR coefficient versus the threshold variable to detect the number and locations of the thresholds. Hansen (1996) considered inference and testing for linearity in situations when a nuisance parameter[1] is not identified under the null hypothesis. He provided a general framework using weighted average and supremum LM tests and gave the asymptotic theory for inference. Hansen (1999a) suggested a sequential testing approach to the regime selection problem. This meant starting with a linear model and adding thresholds until the first acceptance of a null hypothesis. A statistical complication is that the parameters of the TAR model are only identified under the alternative, that is, when the larger model is true. He suggested a likelihood ratio-type test and showed how inference can be conducted using an empirical null distribution of the test statistic generated by the bootstrap. We shall investigate how such a sequential procedure works in practice.

More recently, Gonzalo and Pitarakis (2002), henceforth GP, suggested choosing the number of regimes or thresholds sequentially starting from a linear model (a single regime) and using model selection criteria for choosing between models with m and $m+1$ thresholds. Their argument was that the procedure is easy to use, and as opposed to statistical tests, there is no need to choose significance levels. The work of Gonzalo and Pitarakis was inspired by the results in Bai (1997)

[1] The threshold parameters constitute the nuisance parameters in the TAR case.

and Bai and Perron (1998) who showed that one can estimate break-points in a multiple break model consistently even when the number of breaks estimated is smaller than the actual number of breaks.

Applying model selection criteria or sequential likelihood ratio testing to the present problem requires estimation of models with both m and $m+1$ thresholds. This may not be considered desirable because the larger model is not identified when the smaller model is true. Another potential difficulty with the approach based on information criterion is that implied significance level of the test of testing the model with m against one with $m+1$ thresholds (a comparison with two nested models using a model selection criterion is equivalent to a likelihood ratio test) may vary substantially with the size of the smaller model. On the other hand, the user of sequential likelihood ratio tests is, at least in theory, in full control of the significance level of each test in the sequence. A potential disadvantage of Hansen's tests compared to the GP approach is that they require a substantial computational effort. Besides, GP argue that it is not clear whether or not the sequential approach using these tests can be extended to models with more than two regimes. Some simulation results in this paper illustrate this concern.

The purpose of this paper is to propose a sequential model selection procedure consisting of a sequence of misspecification tests in which a model with m thresholds is tested against one with $m+1$ thresholds. Important features of this method are that standard statistical inference is used in the sequential selection of the number of thresholds and that the modeller has a reasonable if not full control of the significance level of each test. If the true model is a switching regression or a threshold autoregressive one, no claims about asymptotic properties of our tests can be made. Nevertheless, we do claim to have an approximate idea of what the significance levels of the tests in finite samples are. Assuming that the switching regression or threshold autoregressive model under consideration has a fixed number of thresholds, the model selection problem at hand is a finite-sample problem. Therefore, it is sufficient to require that the procedure works in a satisfactory fashion in small and moderate samples. Our simulation experiments suggest that this is indeed the case. Another advantage of our procedure is that it is computationally simple and, as a by-product, yields accurate estimates of the threshold parameters of the TAR model. At each stage only the smaller model is estimated, so that the complication of estimating at least one model that is too large is minimized.

The paper is organized as follows. Section 2 provides the motivation for our procedure and contains a brief overview of smooth transition regression (STR) models on which our technique is based. The technique itself is presented in Section 3. Section 4 contains a simulation study in which our procedure is compared both with the approach of Hansen (1999a) and the one in GP. An empirical application based on the sunspots numbers series can be found in Section 5, and Section 6 contains final remarks.

2 Smooth transition regression model

The general idea underlying our procedure is quite old. Goldfeld and Quandt (1972, pp. 263–264; 1973) considered the estimation of parameters in the switching regression model and pointed out that discontinuity of the log-likelihood complicates the estimation. Their suggestion was to replace the sudden switch or threshold by a smooth transition. This removes the discontinuity, and the parameters of the resulting smooth transition regression model can be estimated by conditional maximum likelihood, using an appropriate iterative algorithm.

In this paper we will apply the same idea - approximation of sudden switches by smooth transitions - to the regime selection problem. That allows us to use standard inference in determining the number of regimes in a TAR model.

A classical logistic STR (LSTR) model for y_t is defined as follows:

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_0 + \mathbf{x}_t'\boldsymbol{\beta}_1 G_{1t} + \varepsilon_t\,, \qquad t = 1, \ldots, T, \tag{1}$$

where $\mathbf{x}_t = (1, x_{1t}, x_{2t}, \ldots, x_{kt})' = (1, \widetilde{\mathbf{x}}_t)'$ is a $((k+1)\times 1)$ vector of explanatory variables, $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ are $((k+1)\times 1)$ parameter vectors and $\{\varepsilon_t\}$ is a sequence of independent, identically distributed normal errors with zero mean and variance σ^2. The transition function G_{1t} in (1) is defined as follows:

$$G_{1t} = G_1(s_t; \gamma_1, c_1) = \left(1 + \exp\{-\gamma_1(s_t - c_1)\}\right)^{-1}, \ \gamma_1 > 0. \tag{2}$$

As $\gamma_1 \to \infty$ in (2), the logistic G_{1t} function approaches the indicator function $I[s_t > c_1]$ and the LSTR model becomes a switching regression (SR) or, in the univariate case, a TAR model with two regimes. The parameter c_1 is then the switch or threshold parameter. Thus the STR model (1) with (2) is a reasonable approximation to the SR model when γ_1 is sufficiently large.

Analogously, we can approximate a multiple-threshold model with a Multiple LSTR (MLSTR) model. For example, an MLSTR model with two transitions has

the form

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_0^* + \mathbf{x}_t'\boldsymbol{\beta}_1^* G_{1t} + \mathbf{x}_t'\boldsymbol{\beta}_2^* G_{2t} + \varepsilon_t, \tag{3}$$

where the transition function $G_{2t} = G_2(s_t; \gamma_2, c_2)$ is again defined as in (2). For the purposes of this paper we set $\gamma_1 = \gamma_2 = \gamma$.

To illustrate how MLSTR model (3) mimics the three-regime TAR model, we reparameterize (3) as follows:

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_1(1 - G_{1t}) + \mathbf{x}_t'\boldsymbol{\beta}_2(G_{1t} - G_{2t}) + \mathbf{x}_t'\boldsymbol{\beta}_3 G_{2t} + \varepsilon_t. \tag{4}$$

Letting $\gamma \to \infty$ we get a piecewise linear form. Figure 1 depicts the three regimes created by G_{1t} and G_{2t} in (4), when $\gamma = 200$, $c_1 = 0.3$, $c_2 = 0.6$ and $s_t = t/T$.

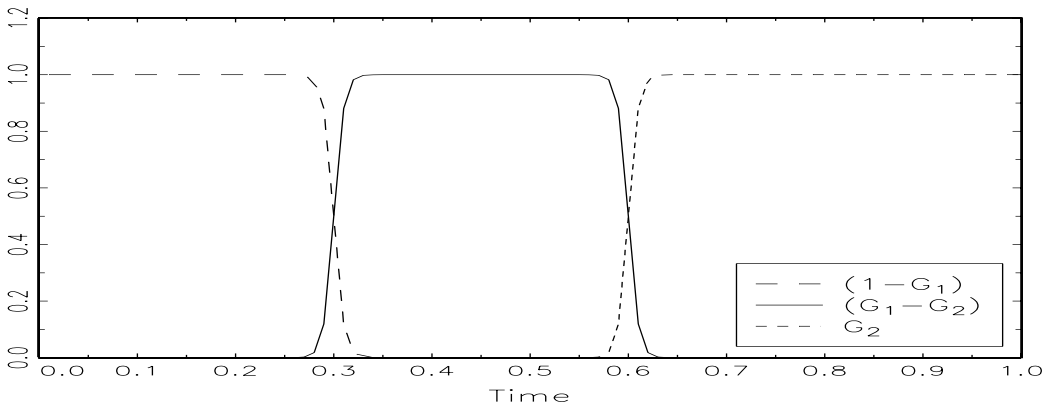

Figure 1: Three regimes

Rearranging the terms in (4) we obtain

$$y_t = \mathbf{x}_t'\boldsymbol{\beta}_1 + \mathbf{x}_t'\left(\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right) G_{1t} + \mathbf{x}_t'\left(\boldsymbol{\beta}_3 - \boldsymbol{\beta}_2\right) G_{2t} + \varepsilon_t, \tag{5}$$

or more generally, in case of $(m+1)$ regimes:

$$\begin{aligned} y_t &= \mathbf{x}_t'\boldsymbol{\beta}_1 + \sum_{j=2}^{m+1} \mathbf{x}_t'\left(\boldsymbol{\beta}_j - \boldsymbol{\beta}_{j-1}\right) G_{j-1,t} + \varepsilon_t \\ &= \mathbf{x}_t'\boldsymbol{\beta}_1\left(1 - G_{1t}\right) + \sum_{j=2}^{m} \mathbf{x}_t'\boldsymbol{\beta}_j\left(G_{j-1,t} - G_{jt}\right) + \mathbf{x}_t'\boldsymbol{\beta}_{m+1} G_{mt} + \varepsilon_t. \end{aligned}$$

Suppose now that the true model is a TAR model with two thresholds. We can approximate this model by the STAR model (5) where γ is large and known. Suppose, however, that we estimate (1) with γ large and known using maximum

likelihood. How does this misspecification affect our threshold parameter estimate? Analogously to GP we argue that underspecification of the number of regimes affects the estimates of slope coefficients $\boldsymbol{\beta}_i$, but hardly those of c_i. In other words, in our three-regime example c_1 can be estimated reasonably accurately even when the number of regimes is misspecified by ignoring G_{2t} in (5).

In the Appendix we show that the average Hessian used as an estimate of the covariance matrix of the average score function, is nearly block-diagonal when γ is large. This means that location parameters can be estimated practically independently of each other, which is necessary for our procedure to work. We also provide simulation evidence from three different three-regime TAR models, showing that when estimating only a two-regime model, the c_1 estimate will be (very close to) one of the true thresholds.

3 Smooth transition approach

In this section we follow GP and consider the univariate TAR model. Our strategy is, however, applicable to switching regression models as well. The starting-point is that the true model is either a linear model or a TAR model (but possibly with just one threshold), so the first choice is between $m = 0$ (linearity) and $m = 1$ (two regimes). As a whole, the procedure works as follows:

1. Test linearity of (1) (i.e $\gamma = 0$ in $G_{1t}(y_{t-d}; \gamma, c_1)$), where $\mathbf{x}_t = (1, \widetilde{\mathbf{x}}_t)' = (1, y_{t-1}, \ldots, y_{t-k})'$. In order to circumvent the identification problem approximate the transition function by its Taylor expansion around $\gamma = 0$. The first-order approximation can be written as $T_1 = \delta_0 + \delta_1 s_t + R_1(\gamma, c; s_t)$, where R_1 is the remainder and δ_0 and δ_1 are constants. Substituting T_1 for G_1 in (1) and reparameterizing yields

$$y_t = \mathbf{x}_t'\boldsymbol{\theta}_0 + (\mathbf{x}_t s_t)'\boldsymbol{\theta}_1 + \varepsilon_t^*, \tag{6}$$

where $\varepsilon_t^* = \varepsilon_t + (\mathbf{x}_t'\boldsymbol{\beta}_1)R_1(\gamma, c; s_t)$. The parameter vector $\boldsymbol{\theta}_1 = \gamma\widetilde{\boldsymbol{\theta}}_1$, where $\widetilde{\boldsymbol{\theta}}_1 \neq \mathbf{0}$, and thus our null hypothesis of linearity in (1) implies $\mathrm{H}_0' : \boldsymbol{\theta}_1 = \mathbf{0}$ in (6). Under $\mathrm{H}_0' : \varepsilon_t^* = \varepsilon_t$. For further discussion of the test, see, for example, Luukkonen, Saikkonen, and Teräsvirta (1988) or Teräsvirta (1998). The resulting test has power against STAR but also against TAR ($\gamma \to \infty$) models. Under the null hypothesis and the assumption $Ey_t^4 < \infty$, the test statistic has an asymptotic χ^2-distribution with $k + 1$ degrees of freedom,

and following the suggestions in earlier papers an F-approximation to it is recommended. The test can be carried out in three stages using just linear regressions:

(a) Regress y_t on $\mathbf{x}_t$ and compute the residual sum of squares $SSR_0 = \frac{1}{T}\sum_{t=1}^{T} \widehat{\varepsilon}_t^2$.

(b) Regress $\widehat{\varepsilon}_t$ (or y_t) on $\mathbf{x}_t$ and $\widetilde{\mathbf{x}}_t s_t$, and compute the residual sum of squares $SSR_1 = \frac{1}{T}\sum_{t=1}^{T} \widehat{v}_t^2$.

(c) Compute

$$F = \frac{(SSR_0 - SSR_1)/k}{SSR_1/(T-2k-1)}$$

that is approximately $F_{k,\,T-2k-1}$ distributed under the null of linearity.

2. If the null hypothesis is rejected at a predetermined significance level α, estimate the parameters of (1) by nonlinear least squares fixing γ at a sufficiently high but finite value. Then the STAR model approximates a TAR model with $m = 1$ and threshold value c_1 while the transition function still retains its smooth character (as a result the likelihood function is well-behaved).

3. If LSTAR model (1) with fixed γ is accepted, test it against a Multiple LSTAR model (5) with transition function G_{2t}. This is done by making use of the first-order Taylor expansion of the transition function G_{2t}, see, for example, Eitrheim and Teräsvirta (1996) or Teräsvirta (1998). Accept (5) if the null hypothesis is rejected at significance level $\tau\alpha$, $0 < \tau < 1$. Reducing the significance level compared to the preceding test favours parsimonious models. Choosing τ is left to the modeller: in the simulations we set $\tau = 1/2$. Starting-values for the estimation may be obtained by using the estimates of $\boldsymbol{\beta}_1, (\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1)$ and c_1. The starting-value for c_2 is obtained by a one-dimensional grid search over a possible set of candidates. This also yields an initial value for $\boldsymbol{\beta}_3$. The estimated model is then tested for another regime. The sequential estimation and testing is continued until the first acceptance of a null hypothesis. This yields the specification for the final model.

4. Estimate consistently the parameters of the final TAR model by conditional least squares (Chan, 1993) or using a dynamic programming algorithm, see Bai and Perron (2003), to estimate the thresholds consistently before estimating the remaining parameters by least squares.

The test can also be constructed using the third-order Taylor approximation of G_{it}. That variant of the test should be more powerful in cases where the process is returning back to its original level after the second threshold, for example. For our STAR-approximation procedure to work in the univariate case we need to assume that ε_t are iid, the transition variable is weakly stationary, and the $2(n+1)$-th moment, where n is the order of Taylor expansion, of y_t exist. It may also be mentioned that the tests can be robustified against heteroskedasticity following Wooldridge (1990).

It should be noted that when a TAR model with m thresholds is tested against one with $m+1$ thresholds, $m \geq 1$, using our test, the asymptotic significance level of the test is unknown. This is the case because the null model is a smooth transition approximation to the null threshold autoregressive model. In testing linearity, however, the asymptotic significance level is known because in that case the null model is not an approximation.

Lack of asymptotic inference may be viewed as a disadvantage, but then, the model selection problem is always a finite-sample problem. Finite sample properties of our technique will be investigated by simulation. The advantages of the STAR-approach are that the tests are computationally simple and that one obtains remarkably accurate values for the threshold parameters even when some of them lie near the smallest or largest observation in the sample.

4 Simulation study

In this section, the small sample performance of the three strategies will be compared by simulation. Choosing between two nested models using an appropriate model selection criterion is equivalent to carrying out the likelihood ratio test, and in some situations the significance level of the model selection criterion based test can be worked out; see, for instance, Teräsvirta and Mellin (1986). In the present case that is not possible even asymptotically because of the identification problem previously mentioned. It is, however, possible to obtain an idea of the empirical size of these tests by simulation. In what follows we shall investigate

both the size of these procedures and their success in finding the correct number of regimes.

In all experiments the true (maximum) lag length of the TAR model is assumed known. In practice one would have to determine the appropriate lag length either simultaneously or before determining the number of regimes. Quite often the lag length is selected prior to building a nonlinear model, using a suitable information criterion.

4.1 Estimating the empirical size

Following GP we simulate univariate autoregressive models, so the alternative to the linear model is a TAR model. We adopt the AR(1) model considered in GP that has the form

$$y_t = \rho y_{t-1} + \varepsilon_t \tag{7}$$

with $\rho = (0.5, 0.7, 0.9, 1.0)$, where $\{\varepsilon_t\} \sim$ nid $(0, 1)$. In order to check the effect of the number of lags on the empirical size of the model selection criteria we also simulate a number of AR(4) models

$$\begin{array}{rcl} \varepsilon_t & = & (1-\rho L)(1-0.2L)(1-0.25L)(1-0.34L)y_t \\ & = & (1-\rho L)(1-0.79L+0.203L^2-0.017L^3)y_t \end{array} \tag{8}$$

with $\rho = (0.5, 0.7, 0.9, 1.0)$, where $\{\varepsilon_t\} \sim$ nid$(0, 1)$ and L is the lag operator. The idea with (8) is to vary the value of the dominant root and, in particular, see what happens when it approaches unity. For $\rho = 1$, the asymptotic distribution theory for testing $\boldsymbol{\theta}_1 = \mathbf{0}$ in (6) is no longer valid.

In the following tables AIC, BIC, BIC2 and BIC3 refer, as in GP, to the model selection criteria

$$Q_T(m) = \max_{c_1,\ldots,c_m} \log\left[\frac{\hat{\sigma}^2}{\hat{\sigma}^2(c_1,\ldots,c_m)}\right] - \frac{\lambda_T}{T} Km \tag{9}$$

with penalty terms $\lambda_T = 2$, $\lambda_T = \log(T)$, $\lambda_T = 2\log(T)$ and $\lambda_T = 3\log(T)$, respectively. In (9), $\hat{\sigma}^2$ is the residual variance in the linear model, $\hat{\sigma}^2(c_1,\ldots,c_m)$ the residual variance of the TAR model with m thresholds, T is the operative number of observations and K is the number of parameters in every regime.

We use, following GP, three different sample sizes $(T = 200, 400, 600)$[2], and

[2]For every sample size we actually generate $T + 200 + k$ observations and discard the first 200 observations from each sample to minimize the impact of starting-values, and use the k extra observations to construct the autoregressive lags of y_t.

	GP				STAR			BOOTSTRAP		
	AIC	BIC	BIC2	BIC3	10%	5%	1%	10%	5%	1%
T=200										
$\rho = 0.5$	80.6	8.75	0.10	0.00	8.75	4.20	0.90	10.00	4.75	0.95
$\rho = 0.7$	81.3	9.15	0.05	0.00	7.65	3.35	0.75	10.10	5.10	1.30
$\rho = 0.9$	80.9	9.50	0.15	0.00	5.00	2.25	0.25	9.60	4.90	0.95
$\rho = 1.0$	88.8	16.35	0.45	1.05	5.65	2.20	0.30	12.30	6.65	1.60
$T = 400$										
$\rho = 0.5$	83.1	4.70	0.00	0.00	9.35	3.85	0.70	9.25	4.20	0.75
$\rho = 0.7$	82.4	5.60	0.05	0.00	8.00	3.65	0.55	9.80	4.95	1.00
$\rho = 0.9$	81.8	6.35	0.00	0.00	6.60	3.20	0.10	10.35	5.05	1.00
$\rho = 1.0$	89.8	9.55	0.20	0.00	5.20	2.45	0.45	10.75	5.15	0.95
$T = 600$										
$\rho = 0.5$	83.8	4.15	0.00	0.00	9.50	4.80	0.85	9.75	4.75	0.85
$\rho = 0.7$	83.3	3.80	0.00	0.00	8.90	4.00	0.70	10.35	4.35	0.75
$\rho = 0.9$	83.8	4.65	0.00	0.00	7.40	3.60	0.40	10.75	5.15	0.90
$\rho = 1.0$	89.4	8.35	0.10	0.00	5.25	2.15	0.25	12.10	5.90	1.40

Table 1: GP-procedure, STAR-approach and Hansen's bootstrap: The empirical size in per cent based on 2000 replications from model (7), using 2000 model-based bootstrap replications in Hansen's procedure.

three different nominal sizes $\alpha = 0.1, 0.05$ and 0.01, respectively. For each DGP and for every sample size, 2000 Monte Carlo replications are carried out.

The results for the AR(1) model (7) appear in Table 1. The threshold or transition variable is assumed to be y_{t-1}. In (7) the intercept is zero, but in practice one would most probably at least tentatively include an intercept in the model. For this reason we assume the intercept to be unknown and a parameter to be estimated from the data. BIC seems to be the only model selection criterion that selects the linear model $4 - 10\%$ of the time, except when $T = 200$ and $\rho = 1$. Both BIC2 and BIC3 point to the correct (linear) model with an empirical probability very close to one and thus have an empirical size close to zero. AIC, as GP also stress, does not work well in this set-up, but its performance is reported here for the sake of comparison.

On the contrary, the STAR-approach[3] has reasonable size properties in the

[3]Throughout Sections 4 and 5 we report the results for test sequences where the test statistics are based on the first-order Taylor approximation. For DGP-s used in this study the discrepancies between the first-order and third-order Taylor approximation approaches were minor.

	GP				STAR			BOOTSTRAP		
	AIC	BIC	BIC2	BIC3	10%	5%	1%	10%	5%	1%
T=200										
$\rho = 0.5$	61.3	0.15	0.00	0.00	8.55	4.25	0.75	9.60	4.30	0.80
$\rho = 0.7$	57.3	0.20	0.00	0.00	7.80	4.05	1.20	11.00	5.40	1.05
$\rho = 0.9$	59.9	0.30	0.00	0.00	6.80	3.10	0.60	10.05	4.55	1.10
$\rho = 1.0$	66.9	0.20	0.05	0.00	7.00	3.55	0.85	11.75	5.95	1.25
$T = 400$										
$\rho = 0.5$	60.7	0.05	0.00	0.00	8.70	4.95	0.85	11.15	5.80	1.35
$\rho = 0.7$	59.9	0.00	0.00	0.00	9.10	4.50	0.90	10.80	5.10	0.85
$\rho = 0.9$	61.8	0.01	0.00	0.00	7.90	3.80	0.85	10.60	5.60	0.95
$\rho = 1.0$	67.6	0.10	0.00	0.00	7.35	3.30	0.60	12.00	6.00	1.35
$T = 600$										
$\rho = 0.5$	62.0	0.05	0.00	0.00	10.00	4.85	0.70	10.30	5.30	1.05
$\rho = 0.7$	62.4	0.05	0.00	0.00	9.60	4.60	0.75	10.35	5.30	1.20
$\rho = 0.9$	62.8	0.00	0.00	0.00	9.25	4.30	0.95	10.85	5.45	1.00
$\rho = 1.0$	69.0	0.00	0.00	0.00	6.55	2.70	0.50	12.05	6.50	1.00

Table 2: GP-procedure, STAR-approach and Hansen's bootstrap: The empirical size in per cent based on 2000 replications from AR(4), using 2000 model-based bootstrap replications.

sense that the empirical sizes are rather close to the ones determined from the F-distribution unless the root of the lag polynomial is close to unity. The linearity test as a whole is seen to be somewhat conservative in small samples. The asymptotic distribution theory of the test is not valid if the AR process is non-stationary, ($\rho = 1$), which explains the increasing size distortion when $\rho \to 1$. Hansen's bootstrap-based test has good size properties already at $T = 200$. Even when the AR process contains a unit root, the empirical size of the test is not too far from the nominal one. The results for the AR(4) model using y_{t-1} as the threshold variable[4] in Table 2 are very different from the ones in Table 1 when the BIC-type model selection criteria are concerned. The increase in the penalty term due to the increased lag length has a remarkable effect on the empirical size. It is practically zero already at $T = 200$. From this we can conjecture that the empirical size of the GP procedure for any AR model with an even longer lag would be practically zero for these criteria at the sample sizes GP considered. AIC is still heavily oversized. The linearity test based on the STAR approximation

[4]The results using any other lag y_{t-d}, $d = 2, 3, 4$, as the threshold variable are very similar to the ones reported here.

tends to be slightly undersized, but at some parameter combinations it competes with Hansen's bootstrap-based method that is well-sized already in small samples.

4.2 Simulating TAR models

In order to consider the performance of the three procedures when the true model is a genuine TAR model we simulate two models also included in the simulation study of GP. One of them has two regimes ($m = 1$) and the other one three ($m = 2$). Furthermore, we complete the experiment with yet another TAR model with $m = 1$. This is done to better demonstrate differences among properties of the three regime-selection procedures. These experiments could be called power simulations except for the fact that the empirical sizes of the three procedures differ substantially from each other.

The error terms in these simulations are constructed to be standard normal variates. We use three different sample sizes ($T = 200, 400, 800$), and three different nominal test size sequences (α, $\alpha\tau$, $\alpha\tau^2, \ldots$), where $\alpha = 0.1, 0.05$ and 0.01, respectively, and $\tau = 1/2$. Our method seems to be robust[5] with respect to the choice of τ. For each DGP and for every sample size, 2000 Monte Carlo replications are carried out.

We begin our STAR-based procedure by testing linearity against (1), assuming the transition variable to be known[6]. If linearity is rejected we proceed to estimate an LSTAR model, fixing the slope parameter $\gamma = 200$. The approach is robust to the choice of γ[7], as long as the logistic function does not deviate much from a step function and the log-likelihood is still well-behaved.

Choosing good starting-values for the optimization algorithm is crucial. We therefore run a grid search over the [.1,.9] interquantile range of the transition variable. This accords with the notion that each regime should contain at least 10% of the total number of observations (see Hansen (1996), Bai and Perron (1998) and GP). After estimating the LSTAR model we look for the second threshold, that is, we test (1) against (5) as discussed in Section 3. If the presence of only a single

[5]We let τ change between $0.1, \ldots, 1.0$. The power loss with respect to the highest-power case was about $0.5 - 1$ percentage points and never greater than 2.8 percentage points (two thresholds, $T = 200$).

[6]It is also possible to define a set of potential transition variables, test against each of them and choose the one giving the strongest rejection (lowest p-value) of linearity.

[7]We let $\gamma = 100, 200, \ldots, 1000$. The largest power loss relative to the maximum, about 2%, occurred at $T = 800, \gamma = 100$. On the average the loss was about 0.6%.

threshold is rejected, we run another grid search to find a good starting-value for the second location parameter, estimate the corresponding MLSTAR model, and proceed until the first acceptance of null hypothesis.

The GP procedure is applied as in the original paper. The required regime size is 10% of the whole sample and thresholds are estimated sequentially, using the Bai (1997) repartition technique. That means re-estimating the threshold parameters conditionally on the initially estimated ones so that each refined estimate is obtained without an underlying neglected regime. In two threshold case, for instance, the first threshold $r^{(1)}$ is re-estimated taking the second threshold estimate $\widehat{r}^{(2)}$ as given and $\widehat{r}^{(2)}$ re-estimated taking the refined estimate of $r^{(1)}$ as given.

When using Hansen's bootstrap-based method we reduce the significance level α as in the STAR-based procedure. Because simulating the likelihood ratio statistics in the sequence can be computationally rather burdensome, we use only 199 model-based bootstrap replications in the application of Hansen's technique. For finding out the power loss that this implies, we refer to Davidson and MacKinnon (2000) who considered the problem of choosing the number of bootstrap replications in bootstrap-based tests. For the test at the 0.10 level the implied power loss should be less than 1%, for a test at 0.01 level the loss should not be greater than $2.5\% - 3\%$.

4.2.1 DGP1: a single threshold model

We begin by considering a TAR model with a single threshold. The data are generated from the following model in GP:

$$y_t = \begin{cases} -3 + 0.5y_{t-1} - 0.9y_{t-2} + \varepsilon_t & y_{t-2} \leq 1.5 \\ 2 + 0.3y_{t-1} + 0.2y_{t-2} + \varepsilon_t & y_{t-2} > 1.5\,. \end{cases} \tag{10}$$

In Table 3 we report the selection frequencies for DGP1 using GP-procedure, i.e. adjusted numbers for Table 6 in GP (page 340). The high frequency for choosing a three-regime model instead of a two-regime model in their original table is due to a slight error in their computer code related to applying the 10% minimum regime size rule mentioned above. The second threshold is often found so close to the first one that there are not sufficiently many observations within the thresholds to make a genuine regime. When the 10% rule is properly applied, the results improve, and in large samples a correct decision is made in over 97% of the occasions.

$T = 200$	$\widehat{m} = 0$	$\widehat{m} = 1$	$\widehat{m} \geq 2$
BIC	7.75	**90.70**	1.55
BIC2	9.60	**90.40**	0.00
BIC3	10.40	**89.60**	0.00
$T = 400$			
BIC	1.30	**97.95**	0.75
BIC2	1.75	**98.25**	0.00
BIC3	2.30	**97.70**	0.00
$T = 800$			
BIC	0.00	**99.70**	0.30
BIC2	0.05	**99.95**	0.00
BIC3	0.05	**99.95**	0.00

Table 3: Adjusted (10% rule applied properly) Table 6 of GP: Selection frequencies for DGP1, $m = 1$.

Results for DGP1 using Hansen's bootstrap and STAR-approach are reported in Table 4. The bootstrap procedure performs about as well as the information criterion based ones. The results for STAR-approach show that the linear model is chosen surprisingly often, about 9% of the time even for $T = 400$.

	STAR			BOOTSTRAP		
$T = 200$	$\widehat{m} = 0$	$\widehat{m} = 1$	$\widehat{m} \geq 2$	$\widehat{m} = 0$	$\widehat{m} = 1$	$\widehat{m} \geq 2$
α =0.10	21.10	**76.30**	2.60	6.80	**86.75**	6.45
α =0.05	22.00	**76.65**	1.35	7.65	**88.55**	3.80
α =0.01	23.50	**76.05**	0.45	8.55	**90.10**	1.35
$T = 400$						
α =0.10	8.65	**89.25**	2.10	1.00	**93.90**	5.10
α =0.05	8.80	**90.25**	0.95	1.15	**96.25**	2.60
α =0.01	9.10	**90.75**	0.15	1.35	**98.05**	0.60
$T = 800$						
α =0.10	1.20	**97.70**	1.10	0.00	**95.90**	4.10
α =0.05	1.20	**98.30**	0.50	0.00	**98.10**	1.90
α =0.01	1.20	**98.75**	0.05	0.05	**99.75**	0.20

Table 4: STAR-approach and Hansen's bootstrap: Selection frequencies for DGP1, $m = 1$.

The reason is that the $\widehat{c}_1$ obtained by numerical optimization sometimes falls outside the $[.1, .9]$ interquantile range and is ignored. Picking a "good" starting-value inside this range does not help when the actual true threshold value lies out

in either of the tails of the empirical density of the threshold variable.

This situation is worth a further comment. GP remarked that DGP1 generates realizations that on the average have approximately the same number of observations in each regime. The true threshold value in our experiment is indeed close to the median of the samples (the average quantile of the threshold value over the replications for any of the three sample sizes is about 0.53). At the same time, in a single sample the true threshold value 1.5 can be very far out in the tails of the empirical distribution, in small samples in particular. Figure 2 shows the frequencies with which the observed deciles of the empirical distribution cover the true threshold value. Decile "0" contains the cases where 1.5 is less than the value of the smallest observation in the sample and decile "11" the cases where the true threshold value exceeds the largest observed value in the sample.

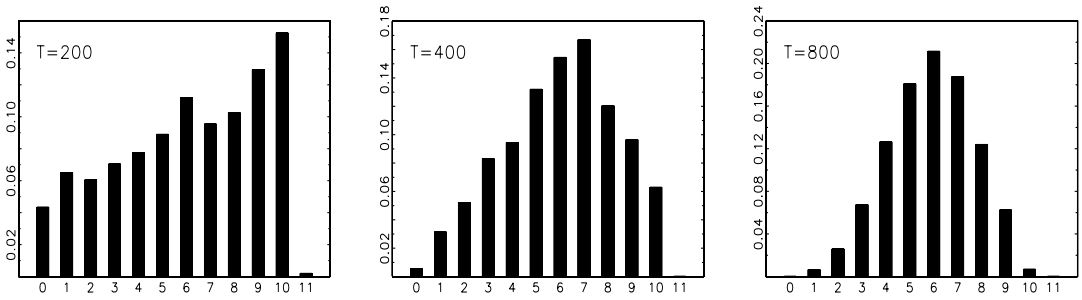

Figure 2: The frequencies with which the observed deciles of the empirical distribution of the threshold variable cover the true threshold value; for $T = 200$, $T = 400$ and $T = 800$.

Consider first the case $T = 200$. In about 4.5% of the realizations the threshold value 1.5 lies outside the range of the simulated series. Thus, at least for these cases a linear model should be selected. In addition to that, in 22% of the cases the true value falls into the first or the last decile. Whenever our location estimate $\widehat{c}_i$ (even if it happens to be close to 1.5) falls outside the $[.1, .9]$ interquantile range the decision has been that it does not signal a genuine threshold. The 10% regime size rule thus explains the high frequencies for selecting $\widehat{m} = 0$ with the STAR-approach. Based on this example we can conjecture that the GP as well as Hansen's procedure might therefore be picking up the second or third best option for the threshold value (from the $[.1, .9]$ range they are restricted to), given that TAR model is preferred to the linear specification. For $T = 400$, the true value is contained in the first decile or is outside the range about 3.5% and in the last one about 6% of the time. The results are quite similar to the previous case, as

the STAR approximation selects the linear model in about 9% of the cases as opposed to 2% for the model selection approach of GP.

The effect of the 10% rule is shown in Table 5 where we report the results of the same experiment after relaxing the regime size restriction for the first threshold. We thus allow its value to belong to the first or the last decile of the observed threshold variable, but we still apply the rule to the next thresholds. Now the linear model is chosen less frequently and the majority of the wrong decisions consists of erroneously detecting a second threshold, except for the smallest sample size. The results are now as good as the ones obtained using Hansen's procedure and signal another advantage of the STAR-approach: the 10% minimum regime size requirement is not necessary when this technique is applied.

$T = 200$	$\widehat{m} = 0$	$\widehat{m} = 1$	$\widehat{m} \geq 2$
$\alpha = 0.10$	8.65	**87.20**	4.15
$\alpha = 0.05$	10.05	**87.80**	2.15
$\alpha = 0.01$	12.05	**87.35**	0.60
$T = 400$			
$\alpha = 0.10$	1.75	**94.20**	4.05
$\alpha = 0.05$	1.95	**96.10**	1.95
$\alpha = 0.01$	2.45	**97.15**	0.40
$T = 800$			
$\alpha = 0.10$	0.00	**96.30**	3.70
$\alpha = 0.05$	0.00	**98.15**	1.85
$\alpha = 0.01$	0.05	**99.45**	0.50

Table 5: STAR-approach: Selection frequencies for DGP1 when not applying the 10% regime-size rule.

We should also mention a difficulty encountered in generating series by the bootstrap for Hansen's procedure. When the optimal threshold value is selected from the [.1, .9] interquantile range and it is not close to the true value, the parameter estimates of the two AR models are (sometimes) far from their true values as well. In that case a number of series generated from the estimated model by bootstrap are explosive. In this experiment, such realizations were discarded and new bootstrap samples generated until the number of valid realizations reached 199. As an example, for sample size $T = 400$, we needed to generate extra bootstrap samples in 5% of the cases. The number of explosive bootstrap series varied between 77 and 2045. We also imposed a "maximum 5000 explosive bootstraps allowed" rule. For DGP1 this rule was flexible enough allowing us to obtain

199 valid bootstrap replicates for every Monte Carlo replication at sample sizes $T = 200, 400$. That was no longer the case for $T = 800$, because it was difficult to generate long non-explosive series. There were 9 cases for which 5000 additional bootstraps were not enough and in the worst case only 24 valid bootstrap series were generated. For these 9 cases the empirical distribution of the F statistic was completed by imputing the missing values with the average of existing bootstrapped statistics.

This difficulty may actually be anticipated. Hansen (1999a, pp. 571), when discussing bootstrapping the distribution for the TAR($m = 1$) vs TAR($m = 2$) test statistic, writes: "We do this with some caution, because there has not yet been a demonstration that a bootstrap procedure can properly approximate the sampling distribution of F_{23} under the SETAR(2) null hypothesis."[8]. In practice, an exploding realization may be taken as a sign of something being wrong with the null model.

4.2.2 DGP2: multiple threshold model

The second DGP, also from GP, is a TAR model with three regimes. It has the following form:

$$y_t = \begin{cases} 2.7 + 0.8y_{t-1} - 0.2y_{t-2} + \varepsilon_t, & y_{t-2} \leq 5 \\ 6 + 1.9y_{t-1} - 1.2y_{t-2} + \varepsilon_t, & 5 < y_{t-2} \leq 12 \\ 1 + 0.7y_{t-1} - 0.3y_{t-2} + \varepsilon_t, & y_{t-2} > 12, \end{cases} \tag{11}$$

where $\{\varepsilon_t\} \sim \text{nid}(0, 1)$. In Table 6 we report the corrected selection frequencies of Table 7 in GP (page 341). Main tendencies are the same as before in that the number of incorrect decisions is small. The thresholds themselves are estimated with reasonable accuracy; see the Appendix.

The results we obtain by applying the STAR-approach to this double-threshold case are quite similar to results from the experiment with one threshold. Power is good even in moderate samples. It does not, however, seem to increase with the sample size. This is due to the fact that the increasing information about the DGP makes the STAR-approximation with a constant γ become less accurate. This disadvantage can be remedied by making the slope parameter γ an increasing function of the sample size.

[8] In Hansen's notation SETAR(n) denotes a model with n regimes, i.e. with $n-1$ thresholds. Furthermore F_{23} denotes the test statistic for testing a 2-regime model against a 3-regime specification i.e. one threshold vs two thresholds.

$T = 200$	$\widehat{m} \leq 1$	$\widehat{m} = 2$	$\widehat{m} \geq 3$
BIC	0.00	**94.20**	5.80
BIC2	0.00	**100.00**	0.00
BIC3	0.00	**100.00**	0.00
$T = 400$			
BIC	0.00	**97.55**	2.45
BIC2	0.00	**100.00**	0.00
BIC3	0.00	**100.00**	0.00
$T = 800$			
BIC	0.00	**98.70**	1.30
BIC2	0.00	**100.00**	0.00
BIC3	0.00	**100.00**	0.00

Table 6: Adjusted (10% rule applied properly) Table 7 of GP.

In this experiment, the problem of explosive realizations when applying the likelihood ratio test sequentially became very severe. When $T = 200$, and when two regimes were tested against three and the "maximum 5000 extra bootstraps" rule was not applied, it took 19252 extra realizations on the average to obtain an empirical distribution based on 199 bootstrap realizations. The maximum number was 490710. The reason for this was that even if one of the thresholds was estimated consistently, merging the two other regimes of the DGP into one (the null model in Hansen's model-based bootstrap) very often led to a highly explosive two-regime model.

The results for $T = 200$ can be found in Table 7. It appears that the sequential likelihood ratio test procedure does not perform as well as the STAR-approach. Simulating the procedure for $T > 200$ is out of the question because of the amount of computations needed to obtain sufficiently many non-explosive realizations. As a whole, one may conclude that the sequential likelihood ratio test procedure may run into problems when the data have been generated by a TAR model with more than two regimes. They can be avoided by making use of the STAR-approximation to the TAR model.

	STAR			BOOTSTRAP		
$T = 200$	$\widehat{m} \leq 1$	$\widehat{m} = 2$	$\widehat{m} \geq 3$	$\widehat{m} \leq 1$	$\widehat{m} = 2$	$\widehat{m} \geq 3$
$\alpha = 0.10$	0.40	**97.80**	1.80	0.00	**63.80**	36.20
$\alpha = 0.05$	0.90	**98.05**	1.05	0.00	**73.80**	26.20
$\alpha = 0.01$	4.90	**94.95**	0.15	0.00	**88.95**	11.05
$T = 400$						
$\alpha = 0.10$	0.00	**97.55**	2.45			
$\alpha = 0.05$	0.00	**99.00**	1.00			
$\alpha = 0.01$	0.00	**99.75**	0.25			
$T = 800$						
$\alpha = 0.10$	0.00	**97.40**	2.60			
$\alpha = 0.05$	0.00	**98.45**	1.55			
$\alpha = 0.01$	0.00	**99.75**	0.25			

Table 7: Selection frequencies for DGP2, STAR-approach and Hansen's bootstrap.

4.2.3 A complementary experiment

GP conclude that overall the BIC criterion displays desirable large sample properties and a reasonably good finite sample behavior. They rightly point out, however, that one should interpret any experimental results with caution since the performance of the criterion depends on the data-generating process. In order to emphasize this feature we complement the experiments in GP by a "real-world" one. The observations are generated by the TAR$(2; 10, 2)$ model in Tong (1990, p. 421), estimated for Wolf's sunspot numbers $1700 - 1979$ transformed as in Ghaddar and Tong (1981). The DGP is

$$y_t = \begin{cases} 1.89 + 0.86y_{t-1} + 0.08y_{t-2} - 0.32y_{t-3} + 0.16y_{t-4} \\ -0.21y_{t-5} - 0.0005y_{t-6} + 0.19y_{t-7} - 0.28y_{t-8} + \\ +0.20y_{t-9} + 0.01y_{t-10} + \varepsilon_t \qquad \text{if} \quad y_{t-8} \leq 11.93 \\ 4.53 + 1.41y_{t-1} - 0.78y_{t-2} + \varepsilon_t \qquad \text{if} \quad y_{t-8} > 11.93, \end{cases} \tag{12}$$

where $\{\varepsilon_t\} \sim \text{nid}(0, 3.734)$. The variance is a "pooled variance"; see Tong (1990, p. 421).

In this experiment our starting-point is an AR(10) model, which implies that the alternative model is a TAR model with ten lags in every regime. An interesting question arises: should one after rejecting the null hypothesis against the TAR model with two regimes determine the lag length in them before proceeding

further (see equation (12) where the second regime only contains two lags), but it is not addressed here.

$T = 200$	$\widehat{m} = 0$	$\widehat{m} = 1$	$\widehat{m} \geq 2$
AIC	0.60	**63.00**	36.40
BIC	71.10	**28.90**	0.00
BIC2	100.00	**0.00**	0.00
BIC3	100.00	**0.00**	0.00
$\alpha = 0.10$	26.90	**70.00**	3.10
$\alpha = 0.05$	34.30	**64.75**	0.95
$\alpha = 0.01$	51.75	**48.05**	0.20
$\alpha_H = 0.10$	14.35	**80.90**	4.75
$\alpha_H = 0.05$	17.45	**80.25**	2.30
$\alpha_H = 0.01$	26.50	**73.05**	0.45
$T = 400$			
AIC	0.05	**64.55**	35.40
BIC	18.80	**81.20**	0.00
BIC2	95.80	**4.20**	0.00
BIC3	100.00	**0.00**	0.00
$\alpha = 0.10$	9.15	**87.05**	3.80
$\alpha = 0.05$	10.35	**87.55**	2.55
$\alpha = 0.01$	14.20	**85.30**	0.50
$\alpha_H = 0.10$	1.85	**93.05**	5.10
$\alpha_H = 0.05$	2.65	**95.00**	2.35
$\alpha_H = 0.01$	4.45	**94.95**	0.60
$T = 800$			
AIC	0.00	**65.80**	34.20
BIC	1.60	**98.40**	0.00
BIC2	21.05	**78.95**	0.00
BIC3	91.45	**8.55**	0.00
$\alpha = 0.10$	5.00	**90.70**	4.30
$\alpha = 0.05$	5.05	**92.45**	2.50
$\alpha = 0.01$	5.20	**94.30**	0.50
$\alpha_H = 0.10$	0.00	**94.70**	5.30
$\alpha_H = 0.05$	0.00	**97.45**	2.55
$\alpha_H = 0.01$	0.00	**99.35**	0.65

Table 8: Selection frequencies for model (12), $m = 1$, for four information criterion-based methods, for the STAR-based approach and for the homoskedastic model-based bootstrap (denoted by subscript $_H$), using starting-significance levels $\alpha = 0.10$, $\alpha = 0.05$, $\alpha = 0.01$.

Results for all three procedures can be found in Table 8. As may be expected from the size simulations, the BIC-type criteria BIC2 and BIC3 strongly favour the linear AR(10) model. Even BIC does that unless the sample size is large ($T = 800$). We also report the results when using AIC, for the sake of comparison in such an extreme case. This criterion works better than any BIC for $T = 200$, but a question arises: which one of these criteria should one use and when? It can be concluded that Hansen's procedure is the best one of the three for this DGP. The STAR-approach is less powerful than Hansen's technique for $T \leq 400$ but performs better than the model selection criteria. In this experiment it overestimates the number of regimes less frequently than Hansen's approach.

5 Application

As an empirical example we consider the original time series of Wolf's sunspot numbers from $1700 - 1979$, transformed as in Ghaddar and Tong (1981). The series with 280 observations is depicted in Figure 3 and exhibits asymmetric cyclical behaviour. It is a very clear-cut example of a nonlinear time series.

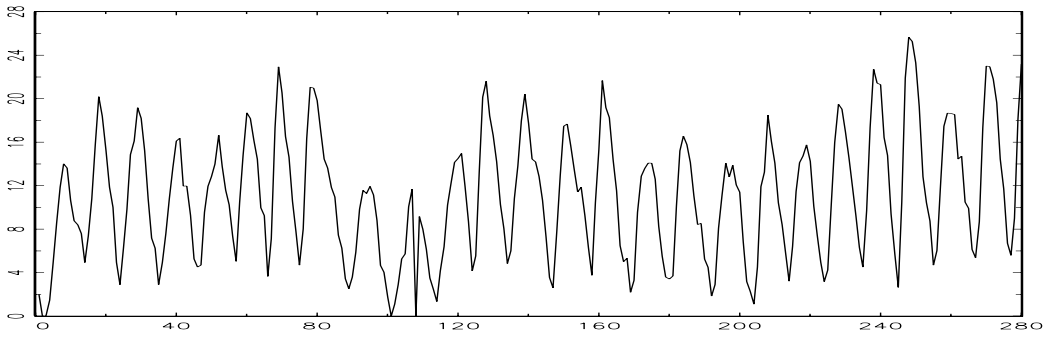

Figure 3: Wolf's sunspot numbers $1700 - 1979$.

When building a TAR model for the series, the autoregressive lag length k for every regime is unknown. It is selected from the linear autoregressive model such that there is no error autocorrelation left in the residuals. We apply the Breusch-Godfrey LM test sequentially: k is increased until the null hypothesis of no error autocorrelation can no longer be rejected at the 5% significance level. This results in $\widehat{k} = 10$.

Using the STAR-approximation we test linearity of the AR(10) model against all ten lags one at a time. Linearity is rejected in eight cases out of ten at 1% level and the lag 8 as the transition variable gives the strongest rejection. From

Table 9 it is seen that the sequential procedure suggested in Section 3 leads to one threshold. Using Hansen's procedure with 2000 bootstrap replications we find one or two thresholds, depending on the initial significance level[9]. To apply the information criterion-based procedure of GP we use delay $d = 8$ found previously. The two information criteria with largest penalty terms, BIC2 and BIC3, prefer the linear model, and only BIC1 is able to detect one threshold.

We also consider lagged first differences of y_t, $\Delta y_{t-d},\ d = 1, \ldots, 10$, as possible threshold variables. All methods choose Δy_{t-1} to be the threshold variable. The results of selecting the number of regimes can be found in Table 9. One threshold is found to be present, with the exception that BIC2 and BIC3 favour a linear model. All estimation methods yield a threshold value close to zero ($\widehat{c}_1 \approx 0.8$), which suggests separate regimes for years with positive and ones with negative growth in sunspot intensity.

	y_{t-d}		Δy_{t-d}	
	$\widehat{d}$	$\widehat{m}$	$\widehat{d}$	$\widehat{m}$
BIC1	8^*	1	1^*	1
BIC2	8^*	0	1^*	0
BIC3	8^*	0	1^*	0
$\alpha = 0.10$	8	1	1	1
$\alpha = 0.05$	8	1	1	1
$\alpha = 0.01$	8	1	1	1
$\alpha_H = 0.10$	8	2	1	1
$\alpha_H = 0.05$	8	2	1	1
$\alpha_H = 0.01$	8	1	1	1

Table 9: Results of sequential model selection procedures. Here $\widehat{d}$ denotes the estimated delay defining the threshold variable and $\widehat{m}$ is the estimated number of thresholds. Asterisk $(*)$ indicates the cases where the threshold variable was assumed known in advance.

[9]From Hansen (1999a) it is known that with a homoskedastic model-based bootstrap one would reject the null of a two-regime model, and with a heteroskedastic bootstrap one would not do that.

6 Final remarks

In this paper we have developed a simple and computationally feasible method for selecting the number of regimes in a switching regression or threshold autoregressive model.

As already pointed out the tests in the STAR-approach can be robustified against heteroskedasticity and thus we only have to assume the independence of errors for the procedure to work. In order to apply Hansen's technique ε_t has to be assumed a uniformly square-integrable martingale difference sequence with respect to the natural filtration, the Borel sigma-field $\mathcal{I}_{t-1} = \sigma(y_{t-1}, y_{t-2}, y_{t-3}, \ldots)$, and $E\varepsilon_t^2 < \infty$. For the bootstrap one also has to assume that the errors are independent. Gonzalo and Pitarakis make quite general assumptions, requiring ε_t to be a real-valued martingale difference sequence with respect to some increasing sequence of sigma fields $\mathcal{F}_t$ generated by $\{(x_{j+1}, z_{j+1}, \varepsilon_j), j \leq t\}$, where z is the threshold variable, and with $E|\varepsilon_t|^{4r} < \infty$ for some $r > 1$. To obtain the limiting distributions of the estimators they make some additional high-level (LLN and FCLT-type) assumptions that exclude integrated processes. GP note that T times the first component in the right-hand side of (9) is the likelihood ratio statistic for testing linearity against a model with m thresholds. Thus their method can, in principle, accommodate the presence of heteroskedasticity through the use of heteroskedasticity-robust versions of this LR or Wald-type statistic. Obviously the method can be generalized such that it will simultaneously allow for selecting the threshold variable as well.

An obvious conclusion from our simulation experiments is that the results of the sequential approach based on model selection criteria are crucially dependent on the number of lags in the TAR model. Admittedly, the users of this approach do not have to choose significance levels for their tests. But then, they face an equivalent problem in the case of GP, which is the one of choosing an appropriate information criterion.

Hansen's bootstrap-based LR-type test can be recommended if it is known that the true number of regimes in the TAR or switching regression model does not exceed two and if computational resources are not a problem. If the existence of more than two regimes cannot be excluded *a priori*, the sequential likelihood ratio test approach may not always work properly. Although the threshold parameters in the model are estimated accurately even when the number of regimes is assumed too small, the estimates of the other parameters in such a model may, due to this

misspecification, cause difficulties when it comes to constructing the empirical distribution of the next test statistic by bootstrap.

The STAR-approach works well in comparison with the other two approaches. It is somewhat conservative, but its performance in selecting the correct TAR model can be deemed acceptable also when the sequential likelihood ratio test procedure excels, that is, when the true model is either linear or has two regimes. The technique is computationally simple, and it performs remarkably well even when a true threshold lies outside the $[.1, .9]$ interquantile range of the observed series. One can relax the minimum regime-size requirement and still estimate the threshold parameters quite accurately.

The discussion in this paper has been restricted to the univariate TAR model, but our technique can also be applied to switching regression models. Besides, it appears that it can be used for determining the number of regimes in the panel threshold regression (PTR) model of Hansen (1999b). This would be done by approximating Hansen's model by the panel smooth transition regression model introduced in González, Teräsvirta, and van Dijk (2004) and using tests described in that paper to determine the number of regimes in the PTR model.

It also seems possible to apply the procedure to detecting the number of breaks in a linear model. This can be in principle done by letting time be the transition variable in the STR model instead of a random transition variable. This possibility is currently being studied by one of the authors.

References

Bai, J. (1997). Estimating multiple breaks one at a time. *Econometric Theory 13*, 315–352.

Bai, J. and P. Perron (1998). Estimating and testing linear models with multiple structural changes. *Econometrica 66*, 47–78.

Bai, J. and P. Perron (2003). Computation and analysis of multiple structural change models. *Journal of Applied Econometrics 18*, 1–22.

Caner, M. and B. E. Hansen (2001). Threshold autoregressions with a unit root. *Econometrica 69*, 1555–1596.

Chan, K. S. (1993). Consistency and limiting distribution of the least squares estimator of a threshold autoregressive model. *Annals of Statistics 21*, 520–533.

Davidson, R. and J. G. MacKinnon (2000). Bootstrap tests: How many bootstraps? *Econometric Reviews 19*, 55–68.

Eitrheim, Ø. and T. Teräsvirta (1996). Testing the adequacy of smooth transition autoregressive models. *Journal of Econometrics 74*, 59–75.

Ghaddar, D. K. and H. Tong (1981). Data transformation and self-exciting threshold autoregression. *Journal of the Royal Statistical Society Ser.C 30*, 238–248.

Goldfeld, S. M. and R. E. Quandt (1972). *Nonlinear Methods in Econometrics*. Amsterdam: North-Holland.

Goldfeld, S. M. and R. E. Quandt (1973). The estimation of structural shifts by switching regressions. *Annals of Economic and Social Measurement 2*, 475–485.

González, A., T. Teräsvirta, and D. van Dijk (2004). Panel smooth transition regression model and an application to investment under credit constraints. Unpublished manuscript, Stockholm School of Economics.

Gonzalo, J. and J.-Y. Pitarakis (2002). Estimation and model selection based inference in single and multiple threshold models. *Journal of Econometrics 110*, 319–352.

Hansen, B. E. (1996). Inference when a nuisance parameter is not identified under the null hypothesis. *Econometrica 64*, 413–430.

Hansen, B. E. (1999a). Testing for linearity. *Journal of Economic Surveys 13*, 551–576.

Hansen, B. E. (1999b). Threshold effects in non-dynamic panels: Estimation, testing and inference. *Journal of Econometrics 93*, 345–368.

Hansen, B. E. (2000). Sample splitting and threshold estimation. *Econometrica 68*, 575–603.

Kapetanios, G. (2003). Threshold models for trended time series. *Empirical Economics 28*, 687–707.

Koop, G. and S. M. Potter (1999). Dynamic asymmetries in U.S unemployment. *Journal of Business and Economic Statistics 17*, 298–312.

Luukkonen, R., P. Saikkonen, and T. Teräsvirta (1988). Testing linearity against smooth transition autoregressive models. *Biometrika 75*, 491–499.

Medeiros, M. C., A. Veiga, and M. Resende (2002). A combinatorial approach to piecewise linear time series analysis. *Journal of Computational and Graphical Statistics 11*, 236–258.

Quandt, R. E. (1958). The estimation of the parameters of a linear regression system obeying two separate regimes. *Journal of the American Statistical Association 53*, 873–880.

Teräsvirta, T. (1998). Modeling economic relationships with smooth transition regression. In A. Ullah and D. E. Giles (Eds.), *Handbook of Applied Economic Statistics*, pp. 507–552. Dekker, New York.

Teräsvirta, T. and I. Mellin (1986). Model selection criteria and model selection tests in regression models. *Scandinavian Journal of Statistics 13*, 159–171.

Tong, H. (1978). On a threshold model. In C. H. Chen (Ed.), *Pattern Recognition and Signal Processing*. Amsterdam: Sijhoff & Noordhoff.

Tong, H. (1990). *Non-linear time series. A dynamical system approach.* Oxford: Oxford University Press.

Tsay, R. S. (1989). Testing and modeling threshold autoregressive processes. *Journal of the American Statistical Association 84*, 231–240.

Wooldridge, J. M. (1990). A unified approach to robust, regression-based specification tests. *Econometric Theory 6*, 17–43.

Appendix A: Properties of threshold estimates

The purpose of this appendix is to give an explanation to the outcome that the threshold parameters can be estimated sequentially with reasonable accuracy from smooth transition approximations to the threshold autoregressive model. Because the STAR model is an approximation to the data-generating process, the arguments are merely suggestive and not based on any asymptotic theory. It suffices to study the block corresponding to the location parameters in the average Hessian and show that it is approximately diagonal. From this it follows that sequential estimation of threshold values yields quite accurate estimates because the estimators of the thresholds are approximately independent. This will be demonstrated using the MLSTAR model (3) that contains two transitions.

Assume that $\{\varepsilon_t\}$, $t = 1, \ldots, T$, is a sequence of identically normally distributed random variables with mean zero and variance σ^2. Then the log-likelihood of the STAR model with two transitions for observation t is

$$l_t = a - \frac{1}{2}\ln\sigma^2 - \frac{\varepsilon_t^2}{2\sigma^2}, \tag{13}$$

where a is a constant, and $\varepsilon_t = y_t - \mathbf{x}_t'\boldsymbol{\beta}_0^* - \mathbf{x}_t'\boldsymbol{\beta}_1^* G_{1t} - \mathbf{x}_t'\boldsymbol{\beta}_2^* G_{2t}$ with $G_{it} = \left(1 + \mathrm{e}^{-\gamma(s_t - c_i)}\right)^{-1}$. Let

$$I(|s_t - c_i| < \varepsilon_\gamma), \quad i = 1, 2, \tag{14}$$

where $I(A)$=1 when A is true and zero otherwise. Thus

$$\begin{aligned} \frac{\partial G_{it}}{\partial c_i} &= \gamma\left(1 + \mathrm{e}^{-\gamma(s_t - c_i)}\right)^{-2} \mathrm{e}^{-\gamma(s_t - c_i)} \\ &= \gamma G_{it}(1 - G_{it}). \end{aligned} \tag{15}$$

For sufficiently large γ, derivative (15) only takes values greater than an arbitrarily small positive constant in a small neighbourhood described by the argument of the indicator function (14). In particular, $\gamma G_{it}(1 - G_{it})|_{s_t = c_i} = \gamma/4$.

Now, assume $|c_1 - c_2| > \delta_\gamma$, where $\delta_\gamma > 0$ is such that if $|s_t - c_1| < \varepsilon_\gamma$, then $|s_t - c_2| > \varepsilon_\gamma$ and vice versa, where $\varepsilon_\gamma > 0$. Setting $L_T = \sum_{t=1}^T l_t$, the elements of the block of interest in the average Hessian are

$$\begin{aligned} \frac{1}{T}\frac{\partial^2 L_T}{\partial c_i^2} &= \frac{1}{\sigma^2}\left(\frac{1}{T}\sum_{t=1}^T (\mathbf{x}_t'\boldsymbol{\beta}_i^*)^2\left(\frac{\partial G_{it}}{\partial c_i}\right)^2 + \frac{1}{T}\sum_{t=1}^T \varepsilon_t(\mathbf{x}_t'\boldsymbol{\beta}_i^*)\frac{\partial^2 G_{it}}{\partial c_i^2}\right) \\ &\approx \frac{1}{\sigma^2}\left(\frac{\gamma^2}{16T}\sum_{t=1}^T (\mathbf{x}_t'\boldsymbol{\beta}_i^*)^2 I(|s_t - c_i| < \varepsilon_\gamma)\right) + o(1), \quad i = 1, 2 \end{aligned} \tag{16}$$

and

$$\begin{aligned}\frac{1}{T}\frac{\partial^2 L_T}{\partial c_1 \partial c_2} &= \frac{1}{\sigma^2}\left(\frac{1}{T}\sum_{t=1}^{T}(\mathbf{x}_t'\boldsymbol{\beta}_1^*)(\mathbf{x}_t'\boldsymbol{\beta}_2^*)\frac{\partial G_{1t}}{\partial c_1}\frac{\partial G_{2t}}{\partial c_2}\right) \\ &\approx \frac{1}{\sigma^2}\left(\frac{\gamma^2}{T}\sum_{t=1}^{T}(\mathbf{x}_t'\boldsymbol{\beta}_1^*)(\mathbf{x}_t'\boldsymbol{\beta}_2^*)I(|s_t-c_1|<\varepsilon_\gamma)I(|s_t-c_2|<\varepsilon_\gamma)\right) \\ &= 0 \end{aligned} \tag{17}$$

because $I(|s_t - c_1| < \varepsilon_\gamma)I(|s_t - c_2| < \varepsilon_\gamma) = 0$. As a consequence, the expression (16) is of larger order of magnitude than (17), and the relevant block of the Hessian is approximately diagonal.

Simulation evidence

To verify that our estimates of c are reasonably accurate when the true number of thresholds is greater than the number of thresholds estimated, consider the DGP2 in our study,

$$y_t = \begin{cases} 2.7 + 0.8y_{t-1} - 0.2y_{t-2} + \varepsilon_t, & y_{t-2} \leq 5 \\ 6 + 1.9y_{t-1} - 1.2y_{t-2} + \varepsilon_t, & 5 < y_{t-2} \leq 12 \\ 1 + 0.7y_{t-1} - 0.3y_{t-2} + \varepsilon_t, & y_{t-2} > 12. \end{cases} \tag{18}$$

When estimating a model with one threshold, the estimates are distributed as follows:

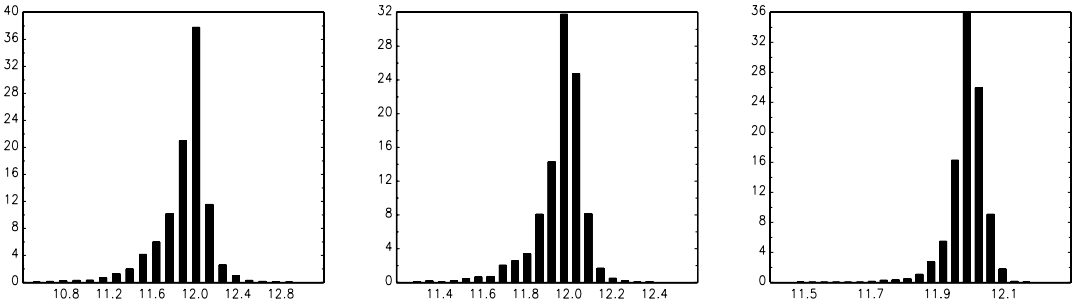

Figure 4: The first threshold estimate distributions for $T = 200$, $T = 400$ and $T = 800$.

The estimates are centered around the true value 12 and the spread of the estimates diminishes when the sample size grows. The same seems to hold for a case where the outer regimes are identical:

$$y_t = \begin{cases} 2.7 + 0.8y_{t-1} - 0.2y_{t-2} + \varepsilon_t, & y_{t-2} \leq 5 \\ 6 + 1.9y_{t-1} - 1.2y_{t-2} + \varepsilon_t, & 5 < y_{t-2} \leq 12 \\ 2.7 + 0.8y_{t-1} - 0.2y_{t-2} + \varepsilon_t, & y_{t-2} > 12. \end{cases} \tag{19}$$

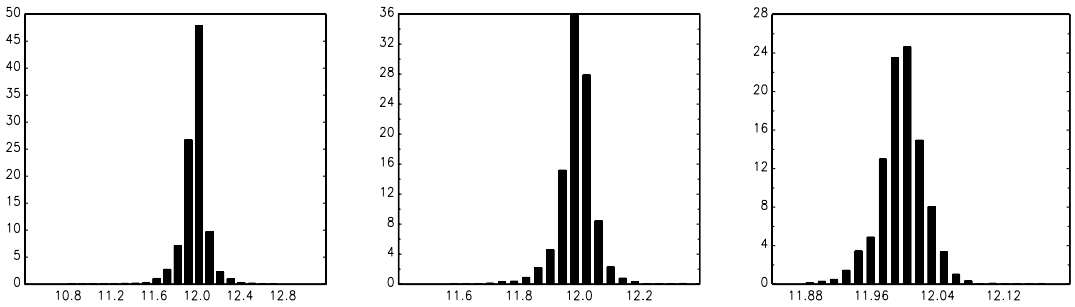

Figure 5: The first threshold estimate distributions for $T = 200$, $T = 400$ and $T = 800$.

Or alternatively:

$$y_t = \begin{cases} 2.7 + 0.8y_{t-1} - 0.2y_{t-2} + \varepsilon_t, & y_{t-2} \leq 3 \\ 1 + 0.7y_{t-1} - 0.3y_{t-2} + \varepsilon_t, & 3 < y_{t-2} \leq 6 \\ 2.7 + 0.8y_{t-1} - 0.2y_{t-2} + \varepsilon_t, & y_{t-2} > 6. \end{cases} \tag{20}$$

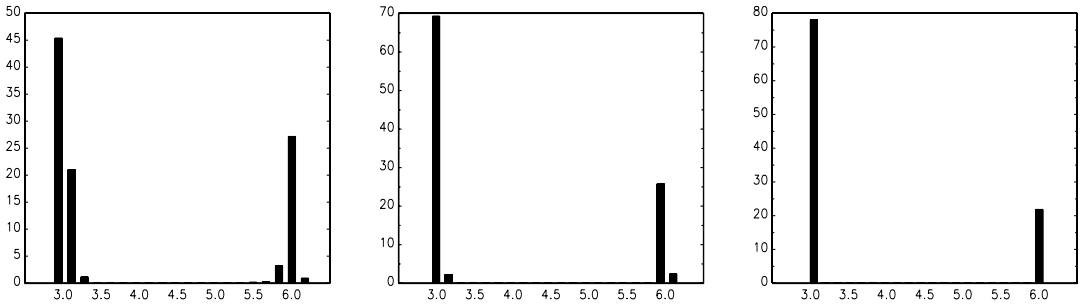

Figure 6: The first threshold estimate distributions for $T = 200$, $T = 400$ and $T = 800$.